A Pre-search Estimation Algorithm for MEDLINE Strategies with Qualifiers

Rulane B. Merz [*] Christopher Cimino, M.D. [†]
G. Octo Barnett, M.D., Dyan Ryan Blewett, John A. Gnassi, M.D.,
Robert Grundmeier and Laurie Hassan

Laboratory of Computer Science
Massachusetts General Hospital
Boston, Massachusetts

ABSTRACT

Inexperienced users of online medical databases often have difficulty formulating their queries. Systems designed to assist them usually do not estimate how effective the initial search strategy will be before performing an actual search. Consequently, the search may find an overwhelming number of citations, or retrieve nothing at all. We have developed an estimation algorithm to predict the outcome of a MEDLINE search. The portion of .the algorithm described here estimates retrieval for strategies containing qualifiers. In test searches, the estimate reduced the trial-and-error of strategy formulation. However, the accuracy of the estimate fell short of expectations. Our results show that pre-search estimation for strategies with qualifiers cannot be performed effectively with only the occurrence data that is presently available. They further imply that automated search intermediaries can benefit from medical knowledge which expresses the relationships that exist between terms.

INTRODUCTION

In their brief summary of the history of databases, Neufeld and Cornog wrote that "databases can almost be said to have created the information industry as we now know it" [11]. Unfortunately, end-users often have difficulty retrieving the information they need. Walker *et al.* [15] studied technical failures of online searches of MEDLINE® , a biomedical database provided by the National Library of Medicine (NLM). They listed several causes of failure: the use of redundant terms or terms that are too general, strategies that are too restrictive, nothing in the database on the topic, and the use of terms that infrequently index a document. Kingsland *et al.* [7] noted that in April 1991, 28% of the searches conducted with the aid of a tool designed to facilitate the use of MEDLINE retrieved no citations; 60% of these were caused by users ANDing terms which were valid, but for which the intersection was null.

We developed QUESTAR (QUery ESTimation And Refinement) to help construct effective initial strategies by predicting the outcome of a search [9, 10]. QUESTAR uses data about the frequency with which terms occur in the database to determine how concepts are related, and thus how often they can be expected to appear together. It obtains the frequency data from the Metathesaurus (Meta), a knowledge source created by the National Library of Medicine's Unified Medical Language System® (UMLS) project [8, 13].

The version of QUESTAR discussed here was designed using Version 1.1 of Meta (Meta-1.1). Currently it is restricted to MEDLINE queries, since Meta-1.1 contained complete occurrence data only for MEDLINE's vocabulary of index terms (Medical Subject Headings, or MeSH).

Previous Work

A system which performs some pre-search estimation is "Animal Welfare Tome.SEARCHER" (AWTS), an intelligent system developed by TOME Associates of London UK [12]. AWTS aids inexperienced users who want to search Agricola, the online agricultural database of the United States Department of Agriculture's National Agricultural Library. It allows a query to be entered as free text, extracts the main concepts, and forms an initial search strategy. It is limited in scope to a single domain (animal welfare), and requires that a dictionary of terms and a classification hierarchy be built.

[*] now at Hughes Information Technology Company. Work performed while a student at Massachusetts Institute of Technology

[†] Albert Einstein College of Medicine

0195-4210/94/$5.00 © 1994 AMIA, Inc.

Chong developed a formula to estimate the percentage of relevant documents that a strategy would retrieve [1]. The formula depended on knowing two things. First, it required the user to prioritize the databases to be searched and the terms in the strategy according to their relevance to the query. Second, it depended on predetermined measures of how much each database and each term would reduce the number of relevant documents that would not be retrieved. These measures were *ad hoc* and did not necessarily reflect what would actually happen. The formula also assumed the independence of concepts in the strategy. Such an assumption has been shown to be unrealistic [5, 14].

Other estimation algorithms have been developed for clustered databases and probabilistic retrieval systems [3, 16, 17]. These algorithms are not directly applicable to keyword-based databases that depend on Boolean retrieval (*e.g.*, MEDLINE). In addition, some of the algorithms require an initial search to be performed so that the frequencies of the terms in the strategy can be determined.

METHODS AND PROCEDURES

QUESTAR's estimates are calculated from Meta's occurrence data for MEDLINE. The occurrence data are of three types. The number of citations which are indexed by a given MeSH heading is the total *frequency of occurrence* of the term in the file. Meta also contains a record of the number of times the term is marked as a MeSH main heading, indicating that it is a main concept in the document. The number of citations which are indexed by a given pair of MeSH main headings is the *frequency of co-occurrence* of the pair. Finally, a MeSH main heading can be combined with *qualifiers*, terms which narrow the focus of the concept described by the main heading. The number of citations indexed by a given main heading/qualifier combination is the frequency of occurrence of the combination.

QUESTAR had reasonable estimates for MeSH headings without co-occurrence data or qualifiers and for MeSH headings with co-occurrence data but without qualifiers [9, 10]. However, it must employ a different method to find the estimate when terms are attached to qualifiers; it must account for the effect a qualifier will have on the retrieval for a concept.

An example will help to explain the estimate for queries with qualifiers. Assume that QUESTAR is given the MeSH main heading/qualifier combinations "zidovudine/therapeutic use" (Z/tu), "AIDS/physiopathology" (A/pp), and "AIDS/drug therapy" (A/dt). The occurrence, co-occurrence and main heading/qualifier occurrence data are given in Tables 1, 2 and 3.

Table 1: Occurrence Data

MeSH main heading	Occurrence Frequency
Zidovudine (Z)	616
AIDS (A)	8291

Table 2: Co-occurrence Data

MeSH main heading pair	Co-occurrence Frequency
Z and A (ZA)	282

Because of the absence of co-occurrence data for pairs of MeSH main heading/qualifier combinations, QUESTAR must use a modified form of the equation given in [9] for queries without co-occurrence data. First, it computes the probability that a combination will appear in an article by dividing the frequency of occurrence of the combination by the total number of articles in the source when the frequency data were collected. It multiplies each of the occurrence probabilities together to obtain the probability that all of the combinations will appear in an article, thereby (falsely) assuming that the occurrences of the combinations are statistically independent. Finally, it multiplies the result by the total number of articles currently in the source to get the expected number of articles that will be retrieved.

For our example, the initial estimate is

$$Estimate = \frac{Z/tu}{T_1} * \frac{(A/pp + A/dt)}{T_1} * T_2 \quad (1)$$

where

T_1 = total number of citations in the source when the occurrence data were tabulated

T_2 = total number of citations currently in the source

Note that if more than one qualifier is attached to the main heading, the frequency of occurrence is the sum of the frequencies of occurrence of the individual main heading/qualifier combinations.

A correction factor is needed, however; the combinations are not independent. If co-occurrence data are available for the MeSH main headings, a correction factor called the *co-occurrence ratio* can be used [10].

$$Ratio = \frac{co\text{-}occurrence(term_1, term_2)}{occurrence(term_1) * occurrence(term_2)} * T_1 \quad (2)$$

The co-occurrence ratio measures the statistical dependence between terms without qualifiers. If it

Table 3: Combination Occurrence Data

MeSH main heading/qualifier	Occurrence Frequency
Zidovudine/therapeutic use (Z/tu)	358
AIDS/physiopathology (A/pp)	142
AIDS/drug therapy (A/dt)	629

is less than one, the two terms are negatively dependent; fewer citations are retrieved than would be expected if the terms were independent. If it is equal to one, the terms are independent, and if it is greater than one, the terms are positively dependent.

MEDLINE had 730,259 citations when the occurrence data in Meta-1.1 were compiled. Using this as the value for T_1, the co-occurrence ratio of the two main headings in our example is 40.3.

To use the co-occurrence ratio, QUESTAR sorts the main headings in ascending order by the magnitude of their ratios. Since the maximum number of articles that will be retrieved by a strategy is bounded by the smallest number of articles that will be retrieved by any pair of terms, QUESTAR begins its calculation of the estimate with the most-negatively/least-positively dependent pair (the pair with the smallest ratio). It multiplies the occurrence probabilities of the terms in the pair combined with their qualifiers to obtain the likelihood that the combinations will appear in the same article. It then multiplies the product by the co-occurrence ratio to correct for its assumption of independence.

Since our example has only two main headings, we simply multiply the estimate by the co-occurrence ratio.

$$Estimate = \frac{Z/tu}{T_1} * \frac{(A/pp + A/dt)}{T_1} * 40.3 * T_2 \quad (3)$$

The co-occurrence ratio is an insufficient measure of the dependencies present in a query with qualifiers. In experiments conducted with the ratio as the only correction for the assumption of independence, the estimate often failed to predict actual retrieval results. Closer inspection of documents retrieved by the queries in conjunction with discussions with physicians identified relationships that exist between qualifiers as an important cause of the discrepancy. A term qualified by "therapeutic use," for example, is more likely to co-occur with a term qualified by "drug therapy" than with a term qualified by "manpower." If more than one term in the query is attached to qualifiers, the dependencies between the qualifiers increase the error in the estimate.

In an effort to quantify the relationships between qualifiers, we computed the co-occurrence ratios for pairs of qualifiers. An online search of MEDLINE was performed to find the occurrence frequencies of the qualifiers and the co-occurrence frequencies of all possible qualifier pairs. The ratio of actual co-occurrence frequency vs. predicted co-occurrence frequency for each pair was then computed using Equation 2 and stored online.

QUESTAR incorporates the additional measure of dependence provided by the qualifier ratios. As previously mentioned, when a term is included in the estimate, the estimate is multiplied by the co-occurrence ratio of which the term is a part. The qualifiers attached to the terms in the co-occurrence ratio are then grouped into pairs consisting of one qualifier from each term. The least-dependent pair of qualifiers, or the pair with the smallest ratio, is found, and the estimate is multiplied by this ratio.

The qualifier ratios in the example are 1.5 and 9 for tu/pp and tu/dt, respectively. Since tu and pp are the least-dependent qualifiers, QUESTAR multiplies the estimate by 1.5. The final estimate is therefore

$$Estimate = \frac{358}{T_1} * \frac{771}{T_1} * 40.3 * 1.5 * T_2 \quad (4)$$

As of January 29, 1993, MEDLINE (1990-93) had 1,051,039 citations. Using this as the value for T_2, substituting in T_1's value and truncating produce a final answer of 32 citations. An actual search retrieved 119 citations.

RESULTS

We collected 24 MEDLINE search strategies containing qualifiers from physicians. QUESTAR computed estimates for these strategies and classified them as either too narrow, acceptable or too broad. The range of acceptable values was 15-30 citations; this was an *ad hoc* definition, based on the need to retrieve at least some useful information, an intuition as to how many citations a busy physician is willing to examine, and our philosophy that it is better to retrieve too much than too little. The estimates were compared with actual search results.

QUESTAR correctly classified 71% of the strategies. Thirteen of the strategies contained a single main heading with a qualifier; 69% of these were classified correctly. Inaccurate predictions may have been due to the sample size; the occurrence frequencies of MeSH main heading/qualifier combinations are small relative to the size of MEDLINE.

Five of the strategies contained multiple main headings, one of which was attached to a qualifier; 80% were classified correctly. However, the accuracy of QUESTAR's predictions was reduced when more than one qualifier was present, even when the estimate included a correction for the relationships

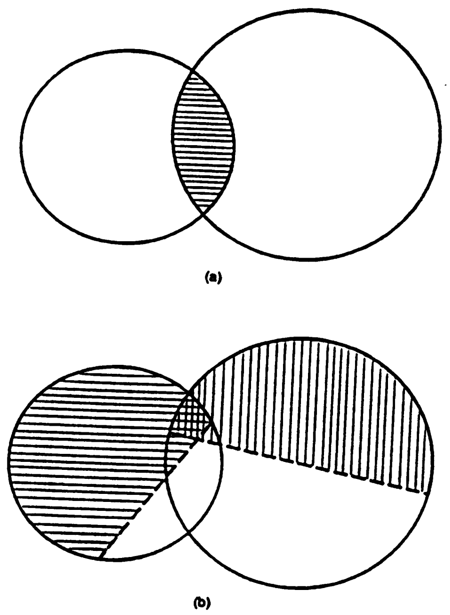

Figure 1: The effects of attaching qualifiers to MeSH main headings. (a) The lined area represents the intersection of two MeSH main headings. (b) The lined areas represent the fraction of the main headings which occur with a particular qualifier. The crosshatched area is the intersection of the MeSH main heading/qualifier combinations.

between qualifiers. Six of the 24 strategies contained more than one main heading/qualifier combination; QUESTAR classified 67% of them correctly.

Discussions with physicians provided an explanation for the drop in accuracy. Not only are qualifiers dependent on each other, but their presence changes the way concepts relate to each other (Fig. 1). For example, the drug "dipyridamole" is not the best medication to prescribe for treating "coronary disease." Adding the qualifier "therapeutic use" to "dipyridamole" and the qualifier "drug therapy" to "coronary disease" decreases the likelihood that the two terms will appear in the same document. Evidence for this conclusion is found in a study performed by Cimino *et al.* [2], which investigated the relationships implied by qualifiers.

DISCUSSION AND CONCLUSION

The estimation algorithm for strategies with qualifiers was not as accurate as we had been led to expect by our previous experiments using strategies without qualifiers. Some inaccuracies are inherent in the use of occurrence data to predict retrieval. Documents stored in MEDLINE reflect the research topics of interest to the medical community at a given time. With new discoveries and the rapid growth of knowledge, the content of the database shifts and changes. Acquired Immunodeficiency Syndrome (AIDS), for example, did not attract much attention until the 1980s, and the term was not added to the MeSH vocabulary until 1983. The occurrence data do not predict these changes, although they do show a history of changes which have already taken place and give some idea of the trends in the literature at the time the data were taken.

Inconsistencies in indexing are also reflected in the occurrence data. Citations on the same subject may be indexed differently, and some citations may be indexed erroneously. In a study of the consistency of MEDLINE's indexing, Funk and Reid [4] found that MeSH main headings, representing concepts that are highly important to the document, are indexed consistently 61.1% of the time. MeSH main headings with qualifiers are indexed consistently only 43.1% of the time. Other studies have shown that the MeSH terms selected vary with the indexer [6]. While QUESTAR knows how many documents have been indexed with a particular term, it cannot determine the accuracy of the indexing. This may lead to reduced or irrelevant retrieval in a search.

In addition to the inaccuracies introduced by the occurrence data, QUESTAR lacks knowledge of the relationship between qualified main headings. Its algorithm assumes that the interdependencies among qualifiers are uniform across MEDLINE; however, the actual relationships depend on the main headings to which the qualifiers are attached. Certain types of medical knowledge would have predictive value, but QUESTAR currently does not contain such knowledge.

Although advances in technology provide access to the information in MEDLINE with no incremental cost, a pre-search estimator can often benefit the user. A search may be time-intensive, especially on small systems with a slow, single-disk CD-ROM; ineffective strategies would create frustration and waste time. An end-user who needs up-to-date information from the definitive source must still perform an online search. Since the cost of an online search is affected by the search strategy and by how much information is retrieved, an algorithm that identifies badly-formulated strategies before a search will reduce the expense. Finally, it is important to avoid overwhelming the naive user with data or providing too little data to be of help. The ability to predict how many articles a strategy will retrieve allows the strategy to be improved before the search, increasing the likelihood that it will retrieve some articles of relevance to the user.

Future Work

QUESTAR already performs some query refinement based on its estimates in an effort to ensure that at least a few citations related to the user's query will be retrieved [10]. The next step would be to incorporate the ability to solicit feedback from the user. The user would not only designate which of the retrieved citations are most relevant to his or her query, but would also choose the most relevant index terms from those citations. QUESTAR could then reformulate the strategy with the chosen terms.

QUESTAR's lack of knowledge about the interactions between concepts introduces error into the estimates. It is not likely or reasonable to expect that complete occurrence data will be available for every possible kind of query. Medical knowledge is more robust and applies to many different kinds of questions. An important area of future work would be to encode the medical information that describes how concepts and qualifiers relate to one another.

Acknowledgments

The authors would like to express their thanks for the technical assistance of Ms. Cindy Schatz, Reference Librarian, Countway Medical Library, Harvard Medical School. This work was supported in part by NLM contract [N01-LM-8-3513] and in part by an educational grant from the Hewlett Packard Corporation. MEDLINE and the Metathesaurus are trademarks of the National Library of Medicine.

References

[1] Chong, H. F. L., *Recall Estimation for Information Retrieval Assistance*, MIT Department of Electrical Engineering and Computer Science, Bachelor's Thesis, 1986.

[2] Cimino, J. J., Mallon, L. J. and Barnett, G. O., "Automated Extraction of Medical Knowledge from Medline Citations", *Proc. of the 12th Annual SCAMC*, Greenes, R. A. (ed.), 1988; 180-184.

[3] Cooper, W. S., Gey, F. C. and Dabney, D. P., "Probabilistic Retrieval Based on Staged Logistic Regression", *Proc. of the 15th International Conference on Research and Development in Information Retrieval*, Belkin, N., Ingwersen, P. and Pejtersen, A. M. (eds.), June 1992; 198-209.

[4] Funk, M. E. and Reid, C. A., "Indexing Consistency in MEDLINE", *Bulletin of the Medical Library Association*, April 1983; 71:176-183.

[5] Harper, D. J. and van Rijsbergen, C. J., "An Evaluation of Feedback in Document Retrieval Using Co-occurrence Data", *J Doc*, September 1978; 34(3):189-216.

[6] Hersh, W. R. and Greenes, R. A., "Information Retrieval in Medicine: State of the Art", *M D Computing*, 1990; 7(5):302-311.

[7] Kingsland, L. C. III, Syed, E. J. and Lindberg, D. A. B., "Coach: An Expert Searcher Program to Assist Grateful Med Users Searching MEDLINE", *MEDINFO 92: Proc. of the 7th World Congress on Medical Informatics*, Lun, K. C., Degoulet, P., Piemme, T. E. and Rienhoff, O. (eds.), September 1992; 382-386.

[8] Lindberg, D. A. B. and Humphreys, B. L., "The UMLS Knowledge Sources: Tools for Building Better User Interfaces", *Proc. of the 14th Annual SCAMC*, Miller, R. A. (ed.), 1990; 121-125.

[9] Merz, R. B., Cimino, C., Barnett, G. O., Blewett, D. R., Gnassi, J. A., Grundmeier, R. and Hassan, L., "Q & A: A Query Formulation Assistant", *Proc. of the 16th Annual SCAMC*, Frisse, M. E. (ed.), 1992; 498-502.

[10] Merz, R. B., *A Pre-Search Estimation Algorithm to Improve Search Strategy Formulation for Medical Databases*, MIT Department of Electrical Engineering and Computer Science, Master's Thesis, 1993.

[11] Neufeld, M. L. and Cornog, M., "Database History: From Dinosaurs to Compact Discs", *JASIS*, July 1986; 37(4):183-190.

[12] "TOME.SEARCHER on Animal Welfare", TOME Associates Ltd., Report and User Guide, June 1990.

[13] "UMLS Metathesaurus", UMLS Fact Sheet, National Library of Medicine, Bethesda, MD, November 1992.

[14] van Rijsbergen, C. J., "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval", *J Doc*, June 1977; 33(2):106-119.

[15] Walker, C. J., McKibbon, K. A., Ryan, N. C., Ramsden, M. F., Fitzgerald, D. and Haynes, R. B., "Methods for Assessing the Competence of Physicians' Use of MEDLINE with GRATEFUL MED", *Proc. of the 13th Annual SCAMC*, Kingsland, L. C. III (ed.), 1989; 441-444.

[16] Yu, C. T., Luk, W. S. and Siu, M. K., "On the Estimation of the Number of Desired Records with Respect to a Given Query", *ACM Trans. on Database Systems*, March 1978; 3(1):41-56.

[17] Yu, C. T., Meng, W. and Park, S., "A Framework for Effective Retrieval", *ACM Trans. on Database Systems*, June 1989; 14(2):147-167.